A Study on Machine Learning Prediction Model for Company Bankruptcy using Features in Time Series Financial Data

Masayoshi Kawamura

Received: 7 December 2021 Accepted: 30 December 2021 Published: 11 January 2022

Abstract

Based on such methods as a discriminant analysis and logistic regression, corporate bankruptcy prediction models have been developed as a means to determine the soundness of a company?s operational status based on its financial statements. However, such analytical methods work with binary variables, and thus, as the only outcome of machine learning, the company in question is considered either likely or unlikely to go bankrupt. However, this is insufficient for business operators who would need to know the possible risk factors of a bankruptcy, allowing them to plan and implement measures to avoid any misfortunes. We have therefore developed a prediction model that not only predicts but also identifies the financial variables that can possibly drive the company to bankruptcy.

Index terms— machine learning; corporate bankruptcy prediction; time-series financial statement data analysis.

1 Introduction

t is extremely important for a company and its stakeholders to have a clear understanding of the operational standing of the company. According to Tasaka [1], to identify corporate credibility based on an analysis of financial statements, studies on "credit analysis" began in the second half of the 19th century, and the Great Depression in the 1930s prompted indepth research into the prediction of bankruptcies in the United States. As stated in Section 2, many bankruptcy prediction models have been proposed in recent years using methods such as a discriminant analysis and logistic regression. However, these analytical methods only return binary outcomes; in most cases, they run machine learning on financial data and predict whether the company in question will or will not go bankrupt. A few existing studies have discussed factors that explain the possible causes of bankruptcy in the given cases. Nevertheless, they either carried out the explanatory consideration manually or employed a different method for the explanatory analysis, falling short of developing a comprehensive (automated) process model. However, from the viewpoint of business operators, knowing those factors that may lead to a bankruptcy is crucial for the preparation of countermeasures. Given this background, we developed a model that facilitates not only a prediction but also the identification of financial variables that may drive the company to bankruptcy.

To evaluate the model, from databases such as kabupro.jp (an online database on listed businesses in Japan), we obtained financial data on financially sound companies and those that went bankrupt. For the operating companies, we referred to the business classification table issued by the Japan Exchange Group, and for each of the 10 primary business categories listed, 10 business entities were randomly selected as the samples. As a result, we verified that the model succeeds in organizing bankrupt companies based on their bankruptcy factors. Furthermore, the model demonstrated its ability to cluster a mixture of sound and bankrupt companies based on their financial patterns, and based on the analyses of financial variables in these clusters, predict specific financial variables that may be exacerbated and lead to bankruptcy.

2 II.

3 Existing and Relevant Studies

Considering Japanese companies, this study deals with a bankruptcy prediction model, and this section presents an overview of existing studies on prediction models targeting businesses within the Japanese context. The new aspect introduced in this study will be described with reference to such studies.

Table 2 lists some of the previous Japanese studies on corporate bankruptcy prediction models. Konoet al. [2] plotted the mean values of their data organized by fiscal year and compared their sample data (of bankrupt companies) with the mean values taken from five listed companies to propose a bankruptcy prediction model.

Okubo [3] proposed a model that evaluates the business management status based on eight patterns of combinations of positive (+) and negative (?) values for the chosen variables, as shown in Table **??**; for example, if the operating cash flow yields a positive value and the investing and financing cash flows yield a negative value, the company in question is in a sound state of business management and will unlikely to go into bankruptcy.

Table **??**: Company performance assessment criteria (source: [3])? ? ? ? ? ? ? ? Operating CF ? ? ? ? - - - - Investing CF ? - ? - ? - ? - Financing CF ? - - ? ? ? - -

Ishikawa et al. [4], Mizoguchi et al. [5], and Jiang [6] employed a discriminant analysis and proposed predictive models for binary bankruptcy/nonbankruptcy outcomes based on machine learning using their respective datasets described in the "Data used" column of Table 2.

In addition, Jidaisho et al. [7] used logistic regression to analyze their data, also shown in Table 2, for machine learning and proposed a prediction model for a binary bankruptcy/non-bankruptcy assessment.

Masuyama [8] also analyzed the financial statements of bankrupt companies by chronologically organizing their data, as described in Table 2. They also drew on surveys administered by the Small Business Institute Japan and management improvement plans of individual companies to compare the actions taken to avoid bankruptcy, based on which they attempted to conduct a bankruptcy prediction.

Finally, Saigo et al. [9] developed a model to evaluate companies by applying the discounted cash flow (DCF) formula to the free cash flow. DCF is a valuation method used to estimate the corporate value at certain discount rates based on the future cash flow expected from a business. Saigo et al. specifically addressed SMEs and micro-businesses and discussed measures to improve their corporate value based on the DCF. For example, they described "cutting unnecessary investments" and "optimizing the equity structure" as financial optimization measures to attain the lowest discount rate, which is one of the components of DCF, and "enhancing the business efficiency" and "investing in profitable businesses (business portfolio optimization)" to maximize the corporate value.

Most of the studies above applied machine learning to their respective financial data and attempted to attain binary outcomes between bankruptcy and nonbankruptcy prediction results. Whereas Masuyama and Jiang both went further and considered those factors responsible for bankruptcy, Masuyama only discussed the factors drawing on some case studies, and the latter employed another method to analyze the detrimental factors after using the discriminant analysis for the binary prediction. They fall short of integrating a series of analyses into a single automated process. Therefore, we developed a model that facilitates not only the prediction but also the identification of financial variables that may contribute to the bankruptcy of a company.

4 Proposed Concept a) Definition of bankrupt company and analyzed data

This study uses the definition of a bankrupt company provided by the Teikoku Databank [10]. We researched bankrupt companies on the Delisting website [11] and obtained the financial statement data of these companies from either kabupro.jp [12] or COSMOS1 [13] (a corporate financial database administered by Teikoku Databank). Regarding the financial statement data of non-bankrupt companies, we referred to the business classification table issued by the Japan Exchange Group and randomly selected 10 companies for each of the 10 primary business categories defined therein. A total of 84 FS data of bankrupt company datasets and 100 FS data of nonbankrupt company datasets were obtained.

Each dataset consists of financial statements of the previous 5 consecutive years, counting from the year of bankruptcy for the 84 bankrupt companies, and from fiscal 2020 for the 100 non-bankrupt companies. Note that, whereas the bankrupt companies were selected to ensure that their corporate sizes and types of trade were unbiased, the same could not be ascertained for nonbankrupt companies because they were randomly selected according to the primary business categories; thus, bias control will be recommended for future evaluative experiments with additional datasets.

5 b) Indexes for bankruptcy prediction

We drew on the data used in the existing studies shown in Table 2, that is, the data in the "Data used" column, based upon which we identified bankruptcy prediction indexes (explanatory variables) for employment in our proposed model. Table 3 lists these indexes, together with the rationale for the choice.

6 Current ratio

This expresses a company's liquid assets against its liabilities due within the current yearperiod, and was chosen because the ratio is considered to decrease as the company nears its bankruptcy. It may be noted that a quick ratio was not selected because the scope of current assets was too narrow.

7 Operating cash flow

This was chosen because it is considered that, in the case of bankruptcy owing to a poor operational performance, the operating cash flow from the main business diminishes.

8 Investing cash flow

This was chosen because the investing cash flow is likely to increase when a company struggles to settle its liabilities, which is attributed to sales of assets such as facilities and company vehicles.

9 Operating cash flow/current liabilities

This indicates a company's ability to settle the liabilities due within the current year from the cash derived from its business activities. This is chosen because the ratio is considered to decrease when the company's performance declines.

10 Inventory turnover (sales revenue/inventory)

A poor performance will lead to a decline in sales revenue, resulting in an increase in inventory (in this case, dead inventory), hence the choice.

11 Operating cash flow/sales revenue

This was chosen because it is possible that bankruptcy may result from a company being excessively short of cash to fulfill its obligations owing to too many illiquid assets such as collectibles despite realizing a large sales revenue.

12 Return on equity (net profit/equity)

Did a bankrupt company practice efficient business management? Was its operating efficiency decreasing over the years prior to the bankruptcy? Knowing the answers to these questions is considered important in formulating preventive measures.

13 Equity/total liabilities

A company likely to go bankrupt undoubtedly has its equity minimized (and in some cases, its liabilities increased), resulting in a decrease in this ratio, and hence the bankruptcy decision. As another reason, it indicates the company's reserve of capital without obligations after offsetting the liabilities.

14 c) Extraction of features from time-series financial data

We will now describe the model used for extracting the features of each of the aforementioned explanatory variables, observed during a 5-year period. Dealing with time-series data, it is common practice to use the logarithm rate of change (logarithmic return) [14,15]. However, this cannot be obtained if a negative value is involved when obtaining a natural logarithm. For this reason, we decided to calculate, instead of the logarithm, the rate of change of the financial indexes over a 5-year period, as shown below: In Formula 1, (Y t) expresses the change rate between the fiscal year t and the preceding year t ?1, and X is each of the eight variable specified as evaluation indexes in the previous section.Y t = (X t ?X t ?1) X t ?1 (1)

We will now discuss a method used to extract the feature values from the trend during the 5-year period (Y t , Y t-1 , ?, Y t-4). The following five patterns (? through ?) were considered using the current asset data included in Table 3, the results of which are shown in Table 4.

? Arithmetic mean of the negative value change rate: This takes as a feature value the mean value of the change rate over the 5-year period (yearequivalent mean value), as in "(Y t-4 + Y t-3 + Y t-2 + Y t-1 +Y t)/5."

? Absolute minimum of negative value change rate: This takes as a feature value the absolute minimum of the negative change rates over the 5-year period, as in "min(|Y t |, |Y t-1 |, ? , |Y t-4 |)."

? Absolute maximum of negative value change rate: This takes as a feature value the absolute maximum of the negative change rates over the 5-year period, as in "max(|Y t |, |Y t-1 |, ? , |Y t-4 |)."

? Sum of negative value change rate: This takes as a feature value the sum of the negative change rates over the 5-year period, as in "sum(Y t , Y t-1 , ? , Y t-4)."

? Year-equivalent change rate between 4 years before and the final year: This takes as a feature value the change rate in years equivalent to between the first and last years of the 5-year period, as in "((Y t -Y t-4)/Y t-4)/4."

The procedure is as follows: Table 4 represents one company that had articulated differences in the rates of change. Here, "? Absolute maximum of negative value change rates" had the largest negative value in the reduction of current assets, and was thus selected as the feature value {FV(Feature Value)}. The formula is expressed as follows, where Y t is obtained, as shown in (1).FV = -max (|Yt|, |Yt-1|, ? , |Yt-4|) , {Yt, Yt-1, ? , Yt-4} < 0 (2)

Here, FV was obtained for each of the eight bankruptcy prediction indices (explanatory variables), which will serve as the input data for clustering in the next section. ? Change rate: Arithmetic mean; ? Absolute minimum of negative change rate; ? Absolute maximum of negative change rate; ? Sum of negative change rate; ? Year-equivalent change rate between 4 years before and the final year.

15 d) Clustering (Machine learning) model

We take the FVs obtained for the eight previously described bankruptcy prediction indexes and create matrix data, as illustrated in Table 5, which will be fed into the machine learning (clustering). Note that the rows are equal to the number of sample datasets, and the names of sampled companies (both bankrupt and non-bankrupt) will appear in the first column. The data in Table 5 are the distance matrix and not the adjacency matrix. The adjacency matrix cannot be used for clustering because distance data are generated between samples (companies). For this reason, we selected a clustering method based on the distance matrix, as advanced by Otsuki [16]. According to this method, the Euclidean distance is obtained based on the principal component scores calculated until the cumulative contribution ratio surpasses 90%, forming a matrix of principal component scores. A silhouette analysis, as shown in (3), is then run on this principal component score matrix, and the number of clusters K is taken at the highest silhouette value at which clustering takes place.?? ?? = ?? ?? ??? ?? ?????? (?? ?? ? ?? ??)**(3)**

In (3), a i is the mean distance between point (node) i and other points in the same cluster, which represents the cluster density; and b i is the smallest mean distance between point i and points in any other cluster than that in which i is a member, representing the dissimilarity to neighboring clusters. This means that, if a cluster partition is applied at the largest mean silhouette value, clustering can be carried out under such conditions in which the clusters are the densest and the most dispersed from one another.

Finally, we describe the modeling and prediction procedures based on this clustering method.

? Modeling procedures: learning datasets (184 companies, with mixed bankruptcy statuses) > normalization > principal component analysis (PCA) > silhouette analysis to determine the number of clusters (K) > clustering with K as the predetermined number of clusters > saving the learning model.

? Prediction procedures: input the dataset non-bankrupt company) > normalization using the learning model > PCA using the learning model > predicting the cluster to which the dataset belongs with reference to the learning model. The prediction outcomes are output for each company, and thus each company will have one prediction result.

16 IV.

17 Clustering Results and Discussions a) Clustering results

We applied the above clustering model to the data described in section 3.1, and as a result, four clusters were formed. Table **??** shows the member distributions. In the next section, the prediction results are discussed. Here, as in a correlation analysis, if thresholds are assumed as a "ratio of bankrupt companies (%) > = 0.7" for a cluster with a high likelihood of bankruptcy and a "ratio of bankrupt companies (%) < = 0.3" for a low likelihood, then a non-bankrupt company is interpreted as not likely to go bankrupt if the company's input dataset belongs to Cluster 1, whereas it is likely an interpretation if the dataset falls within Cluster 3 or 4.

18 Table 6: Member distribution resulting from the clustering b) Corporate bankruptcy prediction using sample datasets

We ran the process ? prediction (test) using sample financial statement data for non-bankrupt companies 1 and 2, the results of which are shown in Figures **??** and 2, respectively. As indicated by the star symbols, Company 1 (Figure **??**) belonged to Cluster 1, whereas Company 2 (Figure **??**) belonged to Cluster 3. Figure 4 shows the member features of Cluster 3, where non-bankrupt Company 2 belongs. The markedly high contributing factor of the bankrupt company group in Cluster 3 is the return on equity (return_on_equity_dmax), followed by the investing cash flow (investment_cash_flow_dmax). Given that the ratio of bankrupt companies in this cluster is 78%, as indicated in Table **??**, it is interpreted that Company 2 can be at risk of bankruptcy if its return on equity and investment in cash flow decline, suggesting the need to consider countermeasures in relation to these factors.

19 Conclusion

By extracting feature values from chronologically ordered financial data used in machine learning, this study sought to develop a prediction model that not only predicts a bankruptcy during a binary judgment, but can also identify the financial variables that are likely to drive a company into bankruptcy. The evaluation test using sample data was successful in that the model clustered bankrupt companies according to the explanatory factors for their bankruptcy. Non-bankrupt companies were also grouped into clusters with the corresponding risk factors of bankruptcy. Thus, the study demonstrated that this model of cluster analysis, based on feature values taken from time-series financial statement data, is effective in predicting and identifying risks of future bankruptcy. [1]

12

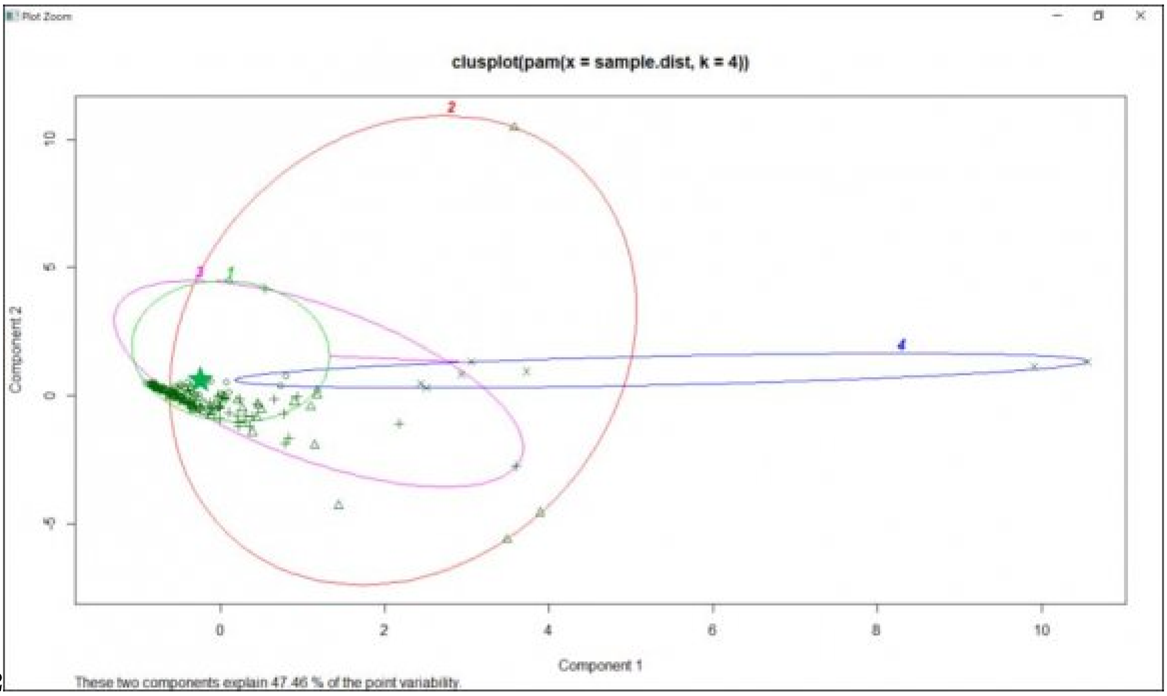

Figure 1: Fig. 1 :Fig. 2 :

[1] © 2022 Global Journals

3

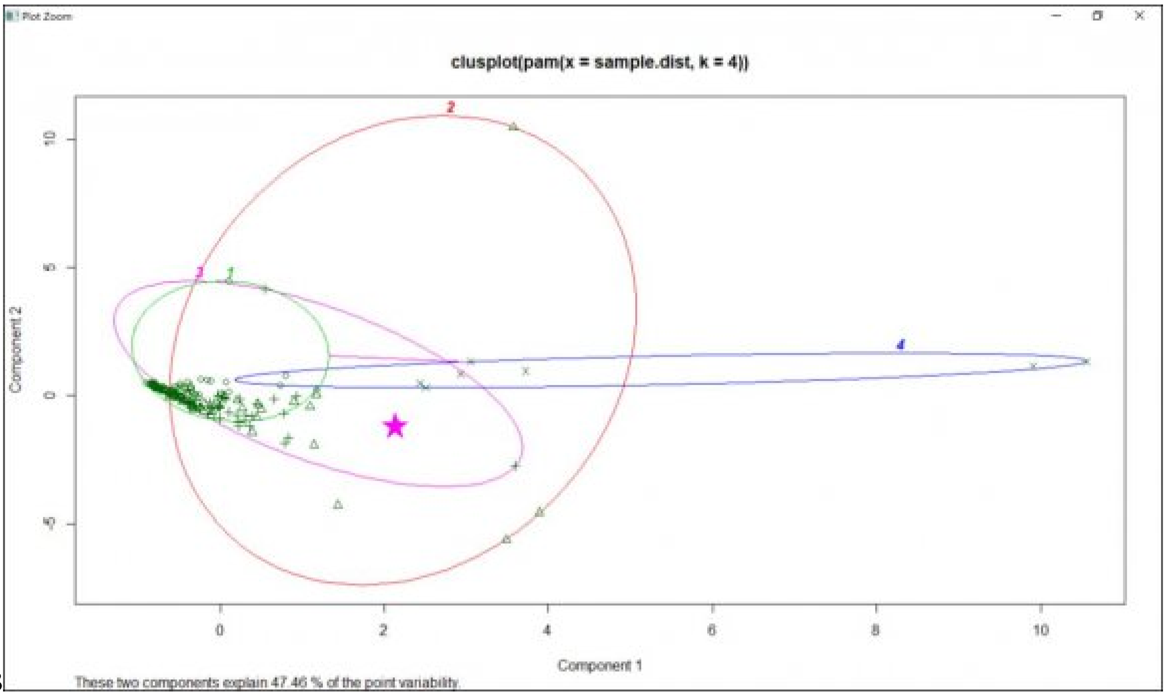

Figure 2: Fig. 3 :

4

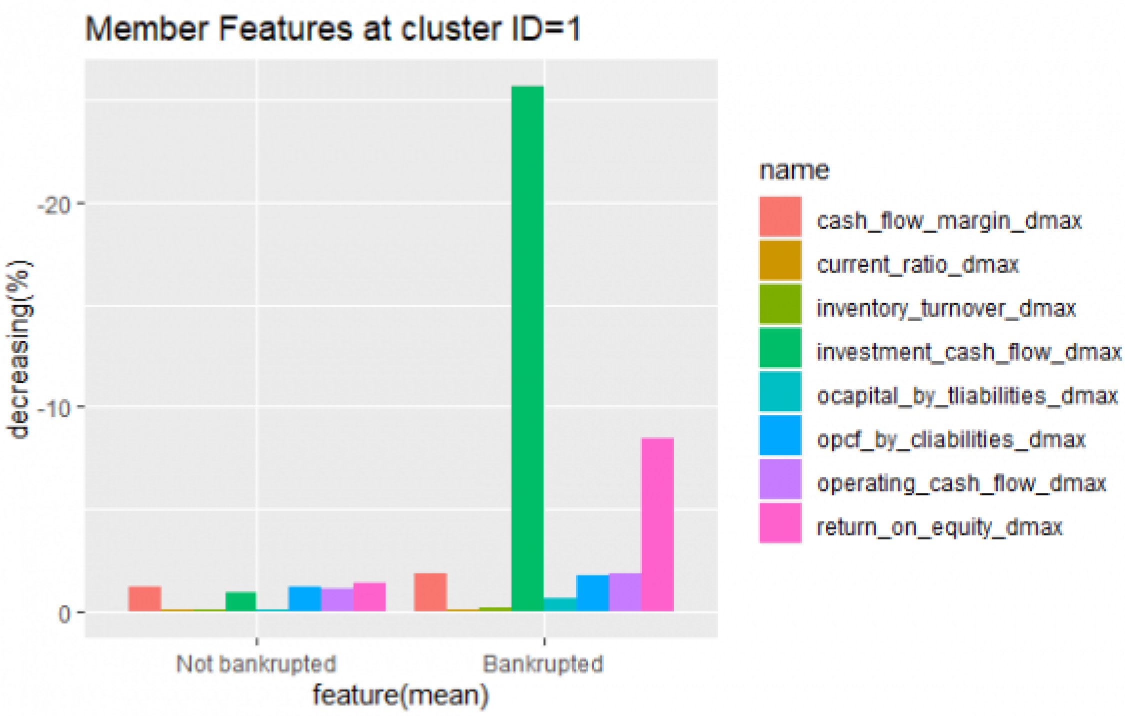

Figure 3: Fig. 4 :

2

Study title	Analytical method	Prediction model
Yosuke Kono et al.: Discussion on the Possibility of Predicting Corporate Bankruptcy [2]	Plotted the mean values of the data organized by fiscal years and compared between the sample data (of bankrupt companies) and the data taken from five listed companies	
Ayaka Okubo: Study on Black-in Bankruptcy Mechanism through Financial Statements Focused on Cash Flow Statement [3]	Developed 8 patterns of corporate cash flows. financial states according to the cor	
Masaru Ishikawa & Ngai Chung Sze: A Study of Corporate Bankruptcies Based on the Cash Flow Information [4]	Discriminant analysis	

[Note: 4. Operating cash flow margin: Operating CF/Sales revenue 5. Corporate CF to sales ratio: (Operating CF + Investing CF)/Sales revenue 6. Total assets to operating CF ratio:]

Figure 4: Table 2 :

3

Explanatory variable	Selection rationale

Figure 5: Table 3 :

4

Current assets

[Note: Note: Key to the numbers in a circle:]

Figure 6: Table 4 :

5

Company	Current ratio	Explanatory variable Operatingcash flow ? ? ?		Equity/Total liabilities
A	FV	FV	FV	FV
B	FV	FV	FV	FV
?	FV	FV	FV	FV
n	FV	FV	FV	FV

Figure 7: Table 5 :

.1 Acknowledgment

This study was funded by the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) through a Grant-in-Aid for Scientific Research (KAKENHI), No. 19K01843.

[Databank (2019)] , Teikoku Databank . `https://www.tdb.co.jp/tosan/teigi.html(accessedon` July 22, 2019.

[Pro Website (2019)] , Kabunushi Pro Website . `http://www.kabupro.jp/code/9963.htm` Oct. 17, 2019.

[Saigo and Nakano ()] *A Basic Study on the Cash Flow Accounting Information and Valuation*, Yasuhiro Saigo , Kazutoyo Nakano . 2012. p. . Bulletin of Toyohashi Sozo University (in Japanese)

[Ishikawa and Chung Sze ()] 'A Study of Corporate Bankruptcies Based on the Cash Flow Information'. Masaru Ishikawa , Ngai Chung Sze . *Toyo Gakuen University Business and Economic Review* 2012. 3 (1) p. . (in Japanese)

[About logarithmic return -Formula for return on stock (2019)] *About logarithmic return -Formula for return on stock*, `http://capitalmarket.jp/post-431/(accessedon` Oct. 17, 2019. (in Japanese)

[Delisting website (2019)] *Delisting website*, `http://delisting.info/index.html` Oct. 17, 2019.

[Mizoguchi and Nakajima ()] *Discriminant analysis of companies and bankruptcy probability estimation*, Ryuji Mizoguchi , Shunsuke Nakajima . 2012. Nanzan University Department of Information Engineering and Science 2012 Dissertations (Abstracts) (in Japanese)

[Kono and Murata ()] 'Discussion on the possibility of predicting corporate bankruptcy'. Yosuke Kono , Masahiko Murata . *Momoyama Gakuin University Gakusei Ronshu*, 2009. (in Japanese)

[Seigotasaka] *Empirical study about a bankruptcy company: It is based on the model of Beaver and Altman*, Seigotasaka . p. . KwanseiGakuin University NII-Electronic Library Service (in Japanese)

[Jidaisho et al. ()] 'Experiment of Risk Prediction on Japanese Companies Using a Bankruptcy Risk Prediction Model'. Koji Jidaisho , Hamido Fujita , Masaki Kurematsu , Jun Hakura . *Iwate Prefectural University Faculty of Software and Information Science Dissertations*, 2017. (in Japanese)

[Feihong ()] 'Financial Forecast and Cash Flow Information'. Jiang Feihong . *Meiji University Studies in Business Administration*, 2003. p. . (in Japanese)

[Otsuki ()] 'Model for Labeling to Latent Factor by Silhouette Clustering Using Principal Component Distance Matrix'. Akira Otsuki . *Journal of Information Processing Society of Japan Database* 2020. 13 (4) p. .

[Okubo ()] Ayaka Okubo . *Study on Black-in Bankruptcy Mechanism through Financial Statements focused on Cash Flow*, 2010. 2010. Junior College of the University of Aizu Research Dissertation Papers for Year (in Japanese)

[Tokuoiwaisako ()] *Serial Correlation of Stock Price Indexes and Cross-autocorrelation of Portfolios by Size*, Tokuoiwaisako . `http://www.ier.hit-u.ac.jp/~iwaisako/research/JRWfinal.pdf(accessedonDec.21` 2018. (in Japanese)

[Masuyama ()] 'The Conditions for Corporate Growth in the Context of Corporate Bankruptcy Analysis'. Yuichi Masuyama . *HikoneRonso Online Journal (Shiga University Economic Journal)* 2017. (413) p. . (in Japanese)